

# A Latent-Variable Bayesian Nonparametric Regression Model

George Karabatsos<sup>1</sup>

*University of Illinois-Chicago, U.S.A.*

and

Stephen G. Walker<sup>2</sup>

*University of Kent, United Kingdom*

January 2, 2012

**Abstract:** We introduce a random partition model for Bayesian nonparametric regression. The model is based on infinitely-many disjoint regions of the range of a latent covariate-dependent Gaussian process. Given a realization of the process, the cluster of dependent variable responses that share a common region are assumed to arise from the same distribution. Also, the latent Gaussian process prior allows for the random partitions (i.e., clusters of the observations) to exhibit dependencies among one another. The model is illustrated through the analysis of a real data set arising from education, and through the analysis of simulated data that were generated from complex data-generating models.

**Keywords:** Bayesian inference; Nonparametric regression; Gaussian process.

---

<sup>1</sup>Corresponding author. Professor, University of Illinois-Chicago, U.S.A., Program in Measurement and Statistics. 1040 W. Harrison St. (MC 147), Chicago, IL 60607. E-mail: gkarabatsos1@gmail.com. Phone: 312-413-1816.

<sup>2</sup>Professor, University of Kent, United Kingdom, School of Mathematics, Statistics & Actuarial Science. Currently, Visiting Professor at The University of Texas at Austin, Division of Statistics and Scientific Computation.

# 1 Introduction

Regression modeling is ubiquitous in many applied research areas. In regression studies, the objective is to estimate specific distributional aspects of a dependent variable  $Y$ , conditional on covariates  $\mathbf{x} = (x_1, \dots, x_p)^\top$  of interest, from a sample data set  $\mathcal{D}_n = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$ , which for notational convenience will be written as  $\mathbf{X}_n = (\mathbf{x}_i^\top)_{i=1}^n$  and  $\mathbf{y} = (y_1, \dots, y_n)^\top$ .

Indeed, for Bayesian nonparametric regression, much research has focused on developing random partition models (RPMs) that follow the general form:

$$\begin{aligned} f(\mathbf{y}|\mathbf{X}_n, \rho_n) &= \prod_{d=1}^{K_n} \prod_{i \in S_d} f(y_i|\mathbf{x}_i, \boldsymbol{\theta}_d) \\ \boldsymbol{\theta}_d &\sim G_0 \\ \rho_n &\sim \pi(\rho_n|\mathbf{X}_n). \end{aligned}$$

In the above,  $\rho_n = \{S_d\}_{d=1}^{K_n}$  denotes a partition of the indices  $\{1, \dots, n\}$  of the sample data  $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$  into  $K_n$  distinct clusters, and  $\pi(\rho_n|\mathbf{X}_n)$  denotes an RPM. These RPM's provide a very broad class of models that encompasses product partition models (PPMs), species sampling models (SSMs), and model-based clustering (MBC); see Quintana (2006) for a review. A PPM is of the form  $\pi(\rho_n|\mathbf{X}_n) = c_0 \prod_{d=1}^{K_n} c(S_d|\mathbf{X}_n)$ , with cohesion functions  $c(S_d|\mathbf{X}_n) \geq 0$  (Hartigan, 1990; Barry & Hartigan, 1993), PPMs have been developed for Bayesian nonparametric regression (Müller & Quintana, 2010; Park & Dunson, 2010; Müller et al., 2011). A SSM assumes the form  $\pi(\rho_n|\mathbf{X}_n) = \pi_{\mathbf{X}_n}(|S_1|, \dots, |S_{K_n}|)$  (Pitman 1996; Ishwaran & James, 2003). The Dirichlet process (Ferguson, 1973), a popular Bayesian nonparametric model, can be characterized either as a special PPM or as a special SSM. On the other hand, with MBC,

a random partition  $\rho_n = \{S_d = \{i : d_i = d\}\}_{d=1}^{K_n}$  is formed by sampling latent indicators  $d_i$ , ( $i = 1, \dots, n$ ), from weights  $\omega_j$  of a discrete mixture model.

In this paper, we develop and illustrate a novel Bayesian nonparametric regression model, which may be characterized as an RPM. The model randomly partitions the  $n$  observations into distinct clusters that each share a common region of a transformed covariate space, and then given the covariate  $\mathbf{x}$ , uses the dependent responses in the covariate region to predict  $Y$ . Specifically, the novel regression model is based on a fixed partition  $(A_j)$  of the range  $\mathbb{R} = \cup_{j=-\infty}^{\infty} (A_j = (j-1, j])$  and a Gaussian Process (GP)  $z(\mathbf{x})$  which induces a random partition by  $\rho_n = \{S_d = \{i : z(\mathbf{x}_i) \in A_d\}\}_{d=1}^{K_n}$ .

To further elaborate, consider the standard Bayesian nonparametric mixture model, with latent variable for the component. Such a model is given by

$$f(y, d) = w_d f(y|\theta_d).$$

The  $d$  classifies which component  $f(y|\theta)$  the observation  $y$  comes from and the weight  $w_d$  is the population probability of coming from component  $d$ . There has been much debate and proposals as to how covariates  $\mathbf{x}$  enter into such a model in a meaningful way. Following the RPM idea it makes most sense that if  $\mathbf{x}$  and  $\mathbf{x}'$  are close then observations  $y$  and  $y'$  would be expected to come from the same component. Hence, it is appropriate to make the  $d$  depend on  $\mathbf{x}$ . A convenient way to achieve this is via a Gaussian process  $z(\mathbf{x})$ , such that

$$d(\mathbf{x}) = j \iff z(\mathbf{x}) \in A_j$$

where  $(A_j)$  is a fixed partition of  $\mathbb{R}$ , i.e.  $\cup_j A_j = \mathbb{R}$  and  $A_j \cap A_{j'} = \emptyset$  for  $j \neq j'$ .

The usual idea of having the weights depend on  $\mathbf{x}$  in the form  $\omega_j(\mathbf{x})$  and having  $\omega_j(\mathbf{x})$  close to  $\omega_j(\mathbf{x}')$  whenever  $\mathbf{x}$  is close to  $\mathbf{x}'$ , is a rather weak condition. While in this case the densities for  $y$  and  $y'$  may be close to each other, there is no suggestion that  $y$  and  $y'$  are coming from the same component, which is the more realistic notion. So what is needed is to have  $y$  close to  $y'$  in probability, rather than simply close in distribution.

Therefore, the proposed model is given by

$$f(y, d|z, \mathbf{x}) = \mathbf{1}(z(\mathbf{x}) \in A_d) f(y|\theta_d).$$

So

$$f(y|z, \mathbf{x}) = \sum_j \mathbf{1}(z(\mathbf{x}) \in A_j) f(y|\theta_j)$$

and

$$f(y|\mathbf{x}) = \sum_j \omega_j(\mathbf{x}) f(y|\theta_j)$$

where

$$\omega_j(\mathbf{x}) = P(z(\mathbf{x}) \in A_j).$$

In Karabatsos and Walker (2012), this model was employed where  $z(\mathbf{x}) \sim \text{n}(\eta(\mathbf{x}), \sigma^2(\mathbf{x}))$ . It was explained in that paper how  $\sigma(\mathbf{x})$  controlled the modes of  $f(y|\mathbf{x})$  and why this was an important aspect of the model in keeping with the idea that  $\mathbf{x}$  close to  $\mathbf{x}'$  determines  $y$  and  $y'$  coming from the same component. That is, for  $\mathbf{x}$  close to  $\mathbf{x}'$ , it can be that  $\omega_j(\mathbf{x})$  and  $\omega_j(\mathbf{x}')$  are both close to 1 for some  $j$ . Henceforth, we refer to Karabatsos and Walker's (2012) model as the "independence model", because it assumes independent latent variables

$z(\mathbf{x}) \sim \mathbf{n}(\eta(\mathbf{x}), \sigma^2(\mathbf{x}))$  and  $z(\mathbf{x}') \sim \mathbf{n}(\eta(\mathbf{x}'), \sigma^2(\mathbf{x}'))$  for any two distinct covariates  $\mathbf{x}$  and  $\mathbf{x}'$ .

In the present paper we acknowledge that it would be further desirable for the  $z$  process to be constructed with dependence; i.e.

$$z(\cdot) \sim \text{GP}[\eta(\cdot), \sigma(\cdot, \cdot)].$$

This will reinforce the notion that it is required for  $\omega_j(\mathbf{x})$  and  $\omega_j(\mathbf{x}')$  to both be close to 1 when  $\mathbf{x}$  is close to  $\mathbf{x}'$ . The dependent Gaussian process facilitates this to a greater extent than under the independent process.

In terms of a RPM, we have

$$P(d_1, \dots, d_n | \mathbf{x}_1, \dots, \mathbf{x}_n) = P(z(\mathbf{x}_1) \in A_{d_1}, \dots, z(\mathbf{x}_n) \in A_{d_n}).$$

This is an appealing version of a probability for the partition as it marginalizes from higher to lower dimensions, addressing the curse of dimensionality. Also, it is clear that since our model allows for the GP to exhibit dependencies among the latent variables  $z(\mathbf{x}_1) \in A_{d_1}, \dots, z(\mathbf{x}_n) \in A_{d_n}$ , it is in a sense more flexible than a PPM because it does not force partitions under the assumption that  $\pi(\rho_n | \mathbf{X}_n)$  is a product prior.

— Insert Figure 1 —

Figure 1 illustrates the mixture weights  $\omega_j(\mathbf{x})$  and the resulting predictive densities  $f(y|\mathbf{x})$  of the model, for a single covariate  $\mathbf{x} = x$  having observed values  $x_1 = 1$ ,  $x_2 = 1.3$ , and  $x_3 = 4$ . Also, the figure assumes  $\eta(x_1) = -.30$ ,  $\eta(x_2) = .21$ ,  $\eta(x_3) = 4.8$ , and the squared-exponential covariance function  $\sigma(x, x') = \sigma_C^2 \exp(-.5||x - x'||^2)$ , and presents the weights

and the densities for small  $\sigma_C^2 = .01$  and for large  $\sigma_C^2 = 10$ . Throughout,  $\|\cdot\|$  denotes the Euclidean norm. As shown, when either  $\sigma_C^2$  is small or large, the mixture weights  $\omega_j(x)$  and the resulting predictive densities  $f(y|x)$  are similar when  $x$  and  $x'$  are close. The weights and densities become more dissimilar as the distance between  $x$  and  $x'$  increases. Also, the parameter  $\sigma_C^2$  controls the number of modes in  $f(y|\mathbf{x})$ . At one extreme, as  $\sigma_C^2$  decreases,  $f(y|\mathbf{x})$  becomes more unimodal. As  $\sigma_C^2$  increases,  $f(y|\mathbf{x})$  becomes more multimodal.

We now describe the layout of the rest of the paper. In Section 2 we fully present our regression model. In Section 3, we illustrate our model through the analysis of real and simulated data sets. In so doing, we compare the predictive performance of our new model, against the previous version of our regression model which assumes independent latent variables  $z(\mathbf{x}_1), \dots, z(\mathbf{x}_n)$ , and against another regression model that is known to provide good predictive performance. Section 4 concludes with a discussion.

## 2 The Regression Model

For a sample set of data  $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ , our Bayesian nonparametric regression model has parameters  $\boldsymbol{\zeta} = ((\boldsymbol{\theta}_j)_{j \in \mathbb{Z}}, \boldsymbol{\beta}, \sigma_C^2, \boldsymbol{\phi})$ , along with latent indicator parameters  $\mathbf{d} = (d_1, \dots, d_n)^\top$ .

The model is defined by:

$$f(\mathbf{y}, \mathbf{d} | \mathbf{X}_n, z, \boldsymbol{\zeta}) = \prod_{i=1}^n f(y_i | \boldsymbol{\theta}_{d_i}) \mathbf{1}(z(\mathbf{x}_i) \in A_{d_i}), \quad i = 1, \dots, n, \quad (1a)$$

$$\pi(\boldsymbol{\theta}) = \prod_{j=-\infty}^{\infty} \pi_j(\boldsymbol{\theta}_j), \quad (1b)$$

$$(z(\mathbf{x}_1), \dots, z(\mathbf{x}_n)) \sim \mathbf{n}_n(\mathbf{X}_{1n} \boldsymbol{\beta}, \sigma_C^2 (\mathcal{C}_\phi(\mathbf{x}_i, \mathbf{x}_l))_{n \times n}), \quad (1c)$$

$$\boldsymbol{\beta}, \sigma_C^2, \boldsymbol{\phi} \sim \mathbf{n}_{p+1}(\boldsymbol{\beta} | \mathbf{m}_\beta, \sigma_C^2 \boldsymbol{\Sigma}_\beta) \text{ga}(\sigma_C^{-2} | a_C, b_C) \pi(\boldsymbol{\phi}) \quad (1d)$$

where  $\mathbf{X}_{1n} = (1, \mathbf{x}_i^\top)_{n \times (p+1)}$ , while  $n_K(\cdot|\cdot, \cdot)$  and  $\text{ga}(\cdot|\cdot, \cdot)$  are respectively the probability density functions of the  $K$ -variate normal and gamma distributions (shape and rate parameterized). As shown, the model is based on a GP, with mean function  $\mathbf{X}_{1n}\boldsymbol{\beta}$  and covariance function matrix  $\sigma_{\mathcal{C}}^2(\mathcal{C}_\phi(\mathbf{x}_i, \mathbf{x}_l))_{n \times n}$ , where  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ , and where  $\mathcal{C}_\phi(\cdot, \cdot)$  is a correlation function that depends on the parameter  $\phi$ .

A standard choice of kernel densities is provided by univariate normal densities  $f(\cdot|\boldsymbol{\theta}_j) = n(\cdot|\mu_j, \sigma_j^2)$  ( $j = 0, \pm 1, \pm 2, \dots$ ), which may be assigned conjugate prior density:

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \prod_{j=-\infty}^{\infty} n(\mu_j|\mu_{\mu j}, \sigma_{\mu j}^2) \text{ga}(\sigma_j^{-2}|\alpha_{\sigma j}, \beta_{\sigma j}).$$

For the covariance function  $\sigma_{\mathcal{C}}^2 \mathcal{C}_\phi(\cdot, \cdot)$ , possible choices of the correlation function include the powered-exponential family  $\mathcal{C}_\phi(\mathbf{x}, \mathbf{x}') = \exp(-\phi_1 \|\mathbf{x} - \mathbf{x}'\|^{\phi_2})$  (for  $\phi_1 > 0$ ;  $0 < \phi_2 \leq 2$ ), the Cauchy family, the Matérn family, as well as families of correlation functions that are either non-stationary or non-isotropic (e.g., Rasmussen & Williams, 2006).

Given data  $\mathcal{D}_n$  likelihood  $\prod_{i=1}^n f(y_i|\mathbf{x}_i; \boldsymbol{\zeta})$ , with  $f(y_i|\mathbf{x}_i; \boldsymbol{\zeta}) = \sum_j f(y|\theta_j)\omega_j(\mathbf{x}; \boldsymbol{\beta}, \sigma_{\mathcal{C}}^2, \phi)$  (see Section 1), and a proper prior density  $\pi(\boldsymbol{\zeta})$  defined over  $\Omega_{\boldsymbol{\zeta}} = \{\boldsymbol{\zeta}\}$ , the posterior density of  $\boldsymbol{\zeta}$  is proper and given by:

$$\pi(\boldsymbol{\zeta}|\mathcal{D}_n) \propto \prod_{i=1}^n f(y_i|\mathbf{x}_i; \boldsymbol{\zeta})\pi(\boldsymbol{\zeta})$$

up to a proportionality constant. Then the posterior predictive density of  $Y$  is defined by:

$$f_n(y|\mathbf{x}) = \int f(y|\mathbf{x}; \boldsymbol{\zeta})\pi(\boldsymbol{\zeta}|\mathcal{D}_n)d\boldsymbol{\zeta},$$

with this density corresponding to posterior predictive mean and variance

$$\mathbb{E}_n(Y_i|\mathbf{x}_i) = \int y f_n(y|\mathbf{x}_i) dy; \quad \text{Var}_n(Y_i|\mathbf{x}_i) = \int \{y - \mathbb{E}(Y_i|\mathbf{x}_i)\}^2 f_n(y|\mathbf{x}_i) dy.$$

In the present paper, in applications of our regression model, our emphasis is in prediction rather than inference of the model parameters  $\boldsymbol{\zeta}$ . Hence, we focus statistical inferences on the posterior predictive density  $f_n(y|\mathbf{x})$ , and functionals of interest.

The posterior densities  $\pi(\boldsymbol{\zeta}|\mathcal{D}_n)$  and  $f_n(y|\mathbf{x})$  can be estimated by using standard Gibbs MCMC sampling methods for infinite-dimensional models, which make use of strategic latent variables (Kalli, Griffin, & Walker, 2010). The Appendix provides more details. Also the Appendix describes how the model and corresponding MCMC methods can be easily extended to handle the analysis of censored observations, discrete dependent variables, and the analysis of spatial or spatio-temporal data via an appropriate modification of the GP covariance function.

## 2.1 Model Assessment and Comparison Methods

After  $M$  regression models are fit to a data set  $\mathcal{D}_n$ , the predictive performance of each model  $m \in \{1, \dots, M\}$  can be assessed by the mean-square predictive-error criterion

$$D(m) = \sum_{i=1}^n \{y_i - \mathbb{E}_n(Y_i|\mathbf{x}_i, m)\}^2 + \sum_{i=1}^n \text{Var}_n(Y_i|\mathbf{x}_i, m) \quad (2)$$

(Gelfand & Ghosh, 1998). The criterion is often used in the practice in the assessment and comparison of Bayesian models (e.g., Gelfand & Banerjee, 2010). The first term of (2)



measures data goodness-of-fit, and the second term is a penalty that is large for models which either over-fit or under-fit the data. The criterion (2) can be re-written as

$$D(m) = \sum_{i=1}^n \int (y - y_i)^2 f_n(y|\mathbf{x}_i, m) dy = \sum_{i=1}^n E_n[(y_i - y)^2 | \mathbf{x}_i, m].$$

So the estimate of  $D(m)$  is obtained by generating posterior predictive samples  $y_i^{\text{pred}(s)} | \mathbf{x}_i$  ( $i = 1, \dots, n$ ) at each iteration  $s = 1, \dots, S$  of the MCMC chain, and then taking

$$\hat{D}(m) = \frac{1}{S} \sum_{s=1}^S \sum_{i=1}^n \left( y_i - y_i^{\text{pred}(s)} \right)^2 = \sum_{i=1}^n \hat{D}_i(m),$$

where the individual quantities  $\hat{D}_i(m)$  can be used to provide a more detailed assessment about a model's predictive performance. The Appendix provides some more details about the MCMC methods for estimating  $D(m)$ .

## 3 Illustrations

### 3.1 *Math Teaching Data*

Here we illustrate the proposed model of equation (1), through the analysis of data that were collected to study a new undergraduate teacher education curriculum, instituted in 2009 by four Chicago-area universities. The study aimed to evaluate the impact of the new curriculum on the ability to teach math among  $n = 89$  of its second-year students. Impact is measured by a dependent variable called "change" (mean=.80; s.d.=.6), which is the change in math teaching ability score of the student, from before (pre-test) and after (post-test)

completing a course in math teaching. Also, there are three covariates. The first covariate is `lmt140`, where `lmt140=1` if the course is learning of math teaching (`lmt`) level 140, and `lmt140=0` if the course is `lmt141` (mean=.73; s.d.=.6). The second covariate is `uic`, which is a 0-1 indicator of whether the student is from the University of Illinois-Chicago, versus one of the other three universities (mean=.60; s.d.=.5). The third covariate is pretest score (mean=-.83; s.d.=.8). Each of the three covariates were z-standardized to have mean zero and variance 1, prior to data analysis.

For the regression model presented in equation (1), we assumed a squared-exponential covariance function for the GP, given by  $\sigma_c^2 \mathcal{C}_\phi(\mathbf{x}, \mathbf{x}') = \sigma_c^2 \exp(-.5 \|\mathbf{x} - \mathbf{x}'\|^2)$ . Also for this model we assigned mostly high-variance priors  $\mu_j \sim_{iid} n(\mu_\mu = 0, \sigma_\mu^2 = 10)$ ,  $\sigma_j^{-2} \sim_{iid} \text{ga}(1, 10^{-3})$ ,  $\beta | \sigma_c^2 \sim n(\mathbf{0}, \sigma_c^2 10^5 \mathbf{I}_{p+1})$ , and  $\sigma_c^{-2} \sim \text{ga}(1, 10^4)$ , to reflect the relative lack of prior information about these parameters. The gamma prior for  $\sigma_c^2$  reflects our prior belief that the conditional density of the change score,  $f(y|\mathbf{x})$ , tends to be unimodal. This implies the belief that the covariates  $\mathbf{x}$  tend to be informative about the dependent variable. To estimate the model, a total of 150,000 sampling iterations of the MCMC algorithm were run (see Appendix), and the last 75,000 samples were used to estimate the model's posterior distribution (after burn-in). The posterior predictive samples and the  $D(m)$  criterion stabilized over the MCMC iterations, and the resulting posterior predictive and  $\hat{D}(m)$  estimates had near-zero 95% Monte Carlo confidence intervals (MCCI) according to a consistent batch means estimator (Jones et al., 2006).

— Insert Figures 2 and 3 —

Figure 2 presents the posterior predictive mean and variance estimates of the change

dependent variable, conditional on the three covariates (lmt140, uic, pretest). The relation between change score and pretest score is quite nonlinear. Figure 3 presents the posterior predictive density estimates of the change score, for a range of values of the pretest scores, while conditioning on uic=1 and lmt140=1. These densities are shown to be skewed and unimodal. In summary, the result show that the new teacher education curriculum tended to have a positive effect on mathematics teaching ability, over time.

We also analyzed the data using a simpler version of our regression model (1), namely the "independence model" (see Section 1), for which we specified  $z(\mathbf{x}_i) \sim_{ind} n((1, \mathbf{x}_i^T)\boldsymbol{\beta}, \sigma_C^2(\mathbf{x}_i))$ , with  $\sigma_C^2(\mathbf{x}) = \exp((1, \mathbf{x}_i^T)\boldsymbol{\lambda})$ , along with priors  $\mu_j \sim_{iid} n(0, 10)$ ,  $\sigma_j^{-2} \sim_{iid} \text{ga}(1, 10^{-3})$ , and  $(\boldsymbol{\beta}, \boldsymbol{\lambda}) \sim n(\mathbf{0}, 10^5 \mathbf{I}_{2(p+1)})$ . Thus, this independence model assumed the same priors for  $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ , as in the GP-based model described earlier. A previous study (Karabatsos & Walker, 2012) showed that the independence model tended to have better predictive performance than 26 other regression models (according to the  $D(m)$  criterion), over many real and simulated data sets, with the BART model (Bayesian Additive Regression Trees model; Chipman, et al. 2010) being among the more competitive models. For the data set under current consideration, the independence model was estimated by 150,000 MCMC sampling iterations, after discarding the first 75,000 samples (burn-in) (see Appendix A of Karabatsos & Walker, 2012). Also, the BART model was fit to the data set, via the generation of 42,000 posterior samples via a Bayesian back-fitting algorithm. For each of these two models, the samples of the  $D(m)$  criterion stabilized over sampling iterations, and the resulting  $\hat{D}(m)$  estimate had a small 95% MCCI.

For the current data set of the  $n = 89$  students under the new teacher education curriculum, our GP-based regression model (1) had a better predictive performance ( $D(m) = 1.3$ ;

MCCI= $\pm 1$ )) than the independence model ( $D(m) = 5.1$ ; MCCI= $\pm 0.8$ )), and the BART model ( $D(m) = 52.5$ ; MCCI= $\pm 0.04$ )) which was fit by implementing the BayesTree package of the R statistical software (Chipman & McCulloch, 2010). Also, for the GP-based model, had no outliers, as the  $\hat{D}_i(m)$  estimates over the  $n = 89$  observations had a 5-number summary (i.e., min, 25%ile, 50%ile, 75%ile, and max) of  $\{.00, .01, .01, .02, .11\}$ . For the independence model, it was  $\{.01, .02, .03, .06, .70\}$ . In summary, it seems that the predictive accuracy of the regression model can be substantially improved by accounting for dependence among the latent variables  $(z(\mathbf{x}_1), \dots, z(\mathbf{x}_n))$ . In the next subsection, we use a simulation study to further investigate this issue.

### 3.2 *Complex Regression Functions*

Here, using a range of complex data-generating models, we conduct a simulation study to compare the predictive performance between the GP-based regression model, the independence model, and the BART model. They include data-generating models where  $f(y|\mathbf{x})$  is a unimodal sampling density, with mean depending on complex functions of  $\mathbf{x}$ . They also include data-generating models where  $f(y|\mathbf{x})$  is a multimodal sampling density, having mean and number of modes that also depend on complex functions of  $\mathbf{x}$ .

For the unimodal  $f(y|\mathbf{x})$  setting, we consider two data-generating models which respectively assumed the following mean functions for the dependent variable:

$$E_1(Y|\mathbf{x}) = 1.9[1.35 + \exp(x_1) \sin(13(x_1 - .6)^2) \exp(-x_2) \sin(7x_2)], \quad (3)$$

$$E_4(Y|\mathbf{x}) = 10 \sin(\pi x_1 x_2) + 20(x_3 - .5)^2 + 10x_4 + 5x_5 + \sum_{k=6}^{10} 0x_k. \quad (4)$$

Equation (3) is a complex 2-dimensional covariate interaction (Hwang, et al., 1994). Equation (4) is a complex function of ten covariates, with 5 covariates irrelevant (Friedman, 1991). With respect to these two functions, we generated a data set of  $n = 225$  observations from  $n(y_i|E_1(Y|\mathbf{x}_i), .0625)u_2(\mathbf{x}_i|0, 1)$ , and we generated another data set of  $n = 100$  observations from  $n(y_i|E_4(Y|\mathbf{x}_i), \sigma_i^2)u_{10}(\mathbf{x}_i|0, 1)$ , for  $i = 1, \dots, n$ , where  $u_K(\mathbf{x}_i|0, 1)$  denotes the density function of a  $K$ -variate uniform distribution.

We simulated two additional data sets under settings where  $f(y|\mathbf{x})$  is multimodal and based on mixtures of normal densities, 10 covariates  $\mathbf{x}$ , and with the number of modes depending on  $\mathbf{x}$ . Specifically, the number of modes  $N_{\text{mod}}(\mathbf{x})$  in the density  $f(y|\mathbf{x})$  ranged from 1 to 4, via the function  $N_{\text{mod}}(\mathbf{x}) = \min(\max(\text{floor}(E_5(Y|\mathbf{x})), 1), 4)$ , with  $E_5(Y|\mathbf{x}) = (3, 1.5, 0, 0, 2, 0, 0, 0)(x_1, \dots, x_8)^\top$ . The four modes are respectively defined by  $E_1(Y|\mathbf{x})$ ,  $E_2(Y|\mathbf{x})$ ,  $E_3(Y|\mathbf{x})$ , and  $E_4(Y|\mathbf{x})$ , with  $E_1(Y|\mathbf{x})$  and  $E_4(Y|\mathbf{x})$  given by (3) and (4), along with

$$E_2(Y|\mathbf{x}) = (-2x_1)^{\mathbb{I}(x_1 < .6)}(-1.2x_1)^{\mathbb{I}(x_1 \geq .6)} + \cos(5\pi x_2)/(1 + 3x_2^2),$$

$$E_3(Y|\mathbf{x}) = ((x_1, x_2, x_3, x_4)\boldsymbol{\beta})^2 \exp((x_1, x_2, x_3, x_4)\boldsymbol{\beta}); \boldsymbol{\beta} = (2, 1, 1, 1)^\top / 7^{1/2}.$$

We simulated a data set of  $n = 100$  observations from a sampling density  $n(y_i|E_{d_i}(Y|\mathbf{x}_i), \sigma_i^2) \times u_{10}(\mathbf{x}_i|0, 1)$ , for  $i = 1, \dots, n$ , with  $d_i \sim u\{1\}$  when  $N_{\text{mod}}(\mathbf{x}_i) = 1$ ,  $d_i \sim u\{1, 2\}$  when  $N_{\text{mod}}(\mathbf{x}_i) = 2$ ,  $d_i \sim u\{1, 2, 3\}$  when  $N_{\text{mod}}(\mathbf{x}_i) = 3$ , and  $d_i \sim u\{1, 2, 3, 4\}$  when  $N_{\text{mod}}(\mathbf{x}_i) = 4$ . Also, we simulated another data set, of  $n = 225$  observations, from the same density.

— Insert Table 1 —

To analyze each of the four simulated data sets described in this subsection, our GP-based

regression model, and our independence model, each assumed priors  $\mu_j \sim_{i.i.d.} n(\hat{\mu}, 100)$ , with  $\hat{\mu}$  the empirical mean of the simulated  $Y$ . Otherwise, each of these models assumed the same priors for their other parameters, and the GP-based model assumed the same squared-exponential covariance function for z-standardized covariates, as in the previous subsection. Moreover, each of these two models were estimated according to 150,000 MCMC sampling iterations, after discarding the first 75,000 samples (burn-in). Also, the BART model was fit to each of the four data sets, via the generation of 300,000 posterior samples. For each of the three models, the  $D(m)$  criterion stabilized over MCMC iterations, and the resulting  $\hat{D}(m)$  estimate yielded a small 95% MCCI.

Table 1 presents the results of the simulation study, in the comparison of the three models, in terms of the mean-squared predictive error  $D(m)$ . The GP-based model obtained the best  $D(m)$  predictive performance for all the simulated data sets. Also, for the GP-based model, the individual  $\hat{D}_i(m)$  predictive errors tended to be quite small, even though the true data-generating models were quite complex. Specifically, for the 2-dimensional unimodal, 10-dimensional unimodal, the multimodal ( $n = 100$ ), and the multimodal ( $n = 225$ ) simulated data sets, the model obtained  $\hat{D}_i(m)$  5-number summaries (i.e., min, 25%ile, 50%ile, 75%ile, and max) of  $\{.01, .01, .02, .02, .15\}$ ,  $\{.01, .02, .03, .04, .10\}$ ,  $\{.00, .02, .02, .03, .07\}$ , and  $\{.01, .02, .02, .04, 4.8\}$ , respectively.

## 4 Conclusions

We have described a Bayesian nonparametric regression model, and demonstrated the suitability of the model through the analysis of both real and simulated data sets. The key

idea of the paper is that close covariates  $x$  and  $x'$  should result in  $y$  and  $y'$  being close in probability, rather than in distribution, which has led to the current prevailing model constructions. Close in probability suggest outcomes from close covariates share a common component distribution which is, in our case, modeled as a normal distribution. For this to happen the weights at a particular component value for these similar covariates should both be close to 1, and to facilitate this a dependent Gaussian process is the most suitable model. Hence, all the aspects of the model play a clear discernible role.

## Appendix: MCMC Algorithm

Our infinite-dimensional regression model can be estimated via the implementation of the MCMC sampling methods of Kalli et al. (2010). This method involves introducing strategic latent variables, to implement exact MCMC algorithms for the estimation of the model's posterior distribution. Specifically, for our regression model (Section 2), we introduce new latent variables  $(u_i)_{i=1}^n$ , and a decreasing function  $\xi_d = \exp(-|d|)$ , such that the model's data likelihood can be rewritten as the joint distribution:

$$\prod_{i=1}^n f(y_i, d_i, u_i | \mathbf{x}_i, z) = \prod_{i=1}^n \{ \mathbf{1}(0 < u_i < \xi_{d_i}) \xi_{d_i}^{-1} f(y_i | \boldsymbol{\theta}_{d_i}) \mathbf{1}(z(\mathbf{x}_i) \in A_{d_i}) \}. \quad (5)$$

Marginalizing over the latent variables  $u_i$  in (5), for each  $i = 1, \dots, n$ , returns the original model (eq. 1a). Thus, given the new latent variables, the infinite-dimensional model can be treated as a finite-dimensional model. This, in turn, permits the use of standard MCMC methods to sample the model's full joint posterior distribution. Given all variables, save the  $(d_i)_{i=1}^n$ , the choice of each  $d_i$  have minimum  $-N_{\max}$  and maximum  $N_{\max}$ , where  $N_{\max} =$

$$\max_i[\max_j \mathbf{1}(u_i < \xi_j)|j|].$$

Then for our regression model, assuming the normal kernel densities  $f(y_i|\boldsymbol{\theta}_j) = \mathbf{n}(y_i|\mu_j, \sigma_j^2)$ ,  $j = 0, \pm 1, \pm 2, \dots$ , each stage of the MCMC algorithm proceeds by sampling from the following full conditional posterior densities:

1.  $\pi(\mu_j|\dots) = \mathbf{n}\left(\mu_j \left| \frac{\mu_{\mu_j}\sigma_j^2 + n_j\sigma_{\mu_j}^2\bar{y}_j}{\sigma_j^2 + n_j\sigma_{\mu_j}^2}, \frac{\sigma_j^2\sigma_{\mu_j}^2}{\sigma_j^2 + n_j\sigma_{\mu_j}^2} \right.\right)$ , for  $j = 0, \pm 1, \dots, \pm N_{\max}$ , with  $n_j = \sum_{i:z_i=j} 1$ ,  $\bar{y}_j = \frac{1}{n_j} \sum_{i:z_i=j} y_i$ ,  $N_{\max} = \max_i[\max_j \mathbf{1}(u_i < \xi_j)|j|]$ , given  $n$  independent uniform random draws  $u_i \sim \mathbf{u}(0, \xi_{|d_i|})$ ,  $i = 1, \dots, n$ ;
2.  $\pi(\sigma_j^{-2}|\dots) = \text{ga}\left(\sigma_j^{-2} \left| \alpha_{\sigma_j} + \frac{1}{2}n_j, \beta_{\sigma_j} + \frac{1}{2} \sum_{i:z_i=j} (y_i - \mu_j)^2 \right.\right)$ , for  $j = 0, \dots, \pm N_{\max}$ ;
3.  $\Pr(d_i = j|\dots) \propto \mathbf{1}(u_i < \xi_j)\xi_j^{-1}\mathbf{n}(y_i|\mu_j, \sigma_j^2)P(z(\mathbf{x}_i) \in A_j)$ , for  $j = 0, \dots, \pm N_{\max}$  and for  $i = 1, \dots, n$ , where  $P(z(\mathbf{x}_i) \in A_j) = \int_{j-1}^j \mathbf{n}(z(\mathbf{x}_i)|\eta_i^*, \psi_{ii}^{-1}) \mathrm{d}z$ , and  $\eta_i^* = (1, \mathbf{x}_i^\top)\boldsymbol{\beta} + \sum_{l \neq i} (-\psi_{il}/\psi_{ii})(z(\mathbf{x}_l) - (1, \mathbf{x}_l^\top)\boldsymbol{\beta})$ , given the precision matrix,  $\Psi_\phi^{(n)} = (\sigma_\phi^2 \mathcal{C}_\phi(\mathbf{x}_i, \mathbf{x}_l))_{n \times n}^{-1} = (\psi_{il})_{n \times n}$ ;
4.  $\pi(z(\mathbf{x}_i)|\dots) \propto \mathbf{1}(z(\mathbf{x}_i) \in A_{d_i} = (d_i - 1, d_i])\mathbf{n}(z(\mathbf{x}_i)|\eta_i^*, \psi_{ii}^{-1})$ , for  $i = 1, \dots, n$ ;
5.  $\pi(\boldsymbol{\beta}|\dots) = \mathbf{n}(\boldsymbol{\beta}|\mathbf{m}_\beta^*, \phi_1 \mathbf{V}_\beta^*)$ , given  $\mathbf{V}_\beta^* = (\mathbf{V}_\beta^{-1} + \mathbf{X}^\top \Psi_\phi^{(n)} \mathbf{X})^{-1}$  and  $\mathbf{m}_\beta^* = \mathbf{V}_\beta^* (\mathbf{V}_\beta^{-1} \mathbf{m}_\beta + \mathbf{X}^\top \Psi_\phi^{(n)} \mathbf{z})$ , where  $\mathbf{z}_n = (z(\mathbf{x}_1), \dots, z(\mathbf{x}_n))^\top$ ;
6.  $\pi(\sigma_\phi^2|\dots) = \text{ga}(\sigma_\phi^{-2} | a_\phi + n/2, b_\phi + \{\mathbf{m}_\beta^\top \mathbf{V}_\beta^{-1} \mathbf{m}_\beta - \mathbf{z}_n^\top \Psi_\phi^{(n)} \mathbf{z}_n - (\mathbf{m}_\beta^*)^\top (\mathbf{V}_\beta^*)^{-1} \mathbf{m}_\beta^*\}/2)$ ;
7.  $\pi(\phi|\dots) \propto \mathbf{n}(z(\mathbf{x}_1), \dots, z(\mathbf{x}_n) | \mathbf{X}_n \boldsymbol{\beta}, \sigma_\phi^2 (\mathcal{C}_\phi(\mathbf{x}_i, \mathbf{x}_l))_{n \times n}) \pi(\phi)$ ;
8.  $f(y^{\text{pred}}|\mathbf{x}, \dots) \propto \mathbf{n}(y|\mu_j, \sigma_j^2) \mathbf{1}(z(\mathbf{x}) \in A_j) \mathbf{n}(z(\mathbf{x})|\mu^*(\mathbf{x}), \sigma_\phi^*(\mathbf{x}))$  for each covariate input  $\mathbf{x}$  of interest, where  $\boldsymbol{\sigma}_\phi(\mathbf{x}) = \sigma_\phi^2 (\mathcal{C}_\phi(\mathbf{x}, \mathbf{x}_1), \dots, \mathcal{C}_\phi(\mathbf{x}, \mathbf{x}_n))^\top$ ,  $\mu^*(\mathbf{x}) = (1, \mathbf{x}^\top)\boldsymbol{\beta} + \boldsymbol{\sigma}_\phi(\mathbf{x})^\top \Psi_\phi^{(n)} (\mathbf{z}_n - \mathbf{X}_{1n} \boldsymbol{\beta})$ , and  $\sigma_\phi^*(\mathbf{x}) = \sigma_\phi^2 \mathcal{C}_\phi(\mathbf{x}, \mathbf{x}) - \boldsymbol{\sigma}_\phi(\mathbf{x})^\top \Psi_\phi^{(n)} \boldsymbol{\sigma}_\phi(\mathbf{x})$ .



The full conditionals in Steps 1-6 and 8 can be sampled directly, using standard theory for Bayesian linear models, GP models, and standard methods for sampling truncated normal distributions (e.g., O’Hagan & Forster, 2004; Damien & Walker, 2001). The full conditional in Step 7 can be sampled using a Metropolis-Hastings or another rejection-sampling algorithm, if necessary. Step 8 of the MCMC algorithm provides samples from the posterior predictive density  $f_n(y|\mathbf{x})$  of the regression model. The full 8-step sampling algorithm is repeated a large number  $S$  of times, to construct a discrete-time Harris ergodic Markov chain  $\{\boldsymbol{\zeta}^{(s)} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\beta}, \sigma_c^2, \boldsymbol{\phi})^{(s)}\}_{s=1}^S$  having a posterior distribution  $\Pi(\boldsymbol{\zeta}|\mathcal{D}_n)$  as its stationary distribution, provided a proper prior  $\Pi(\boldsymbol{\zeta})$ . We have written MATLAB (2012, The MathWorks, Natick, MA) code that implements the MCMC sampling algorithm. Standard methods can be used to check whether the MCMC algorithm has generated a sufficiently-large number of samples from the model’s posterior distribution. Specifically, given MCMC samples  $\{\boldsymbol{\zeta}^{(s)}\}_{s=1}^S$  generated by the algorithm, univariate trace plots of these samples can be used to evaluate the mixing of the chain (i.e., the degree to which the chain explores the support of the posterior distribution). Also, for a posterior (moment or quantile) estimate of any chosen scalar functional  $\varphi(\boldsymbol{\zeta})$ , a Monte Carlo Confidence Interval (MCCI) can be computed via an applications of a batch means method (for posterior moment estimates) or a subsampling method (for posterior quantile estimate) applied to the MCMC samples  $\{\varphi(\boldsymbol{\zeta}^{(s)})\}_{s=1}^S$ .

Simple modifications of the MCMC algorithm and/or our model (Section 2) can be used to address other important data analysis tasks:

- Multiple imputation of a censored dependent response  $y_i$ : At each iteration of the MCMC algorithm, a plausible value of a dependent response  $y_i$  that is censored and known

only to fall within an interval  $(a_{y_i}, b_{y_i}]$ , is sampled from the full conditional posterior predictive density  $\pi(y|\mathbf{x}_i, \dots) \propto n(y|\mu_{d_i}, \sigma_{d_i}^2)\mathbf{1}(y \in (a_{y_i}, b_{y_i}])$ , and then is imputed as the updated value of  $y_i$ .

- Discrete-valued dependent variable: Our regression model can be extended to handle ordinal discrete-valued dependent variable responses,  $y_i \in \{0, 1, \dots, C_i^{\max} \geq 1\}$  ( $i = 1, \dots, n$ ), by using instead probit kernels of the form  $f(c|\boldsymbol{\theta}_j) = \int_{\mathcal{A}(c)} n(y^*|\mu_j, \sigma_j^2) dy^*$  ( $j = 0, \pm 1, \pm 2, \dots$ ), for disjoint sets  $\mathcal{A}(c)$  such that  $\cup_{c=0}^C \mathcal{A}(c) = \mathbb{R}$ . In this case, we would add a step to the existing MCMC algorithm, to sample from the full conditional posterior density of the latent variables  $\pi(y_i^*|\mathbf{x}_i, \dots) \propto n(y_i^*|\mu_{d_i}, \sigma_{d_i}^2)\mathbf{1}(y_i^* \in \mathcal{A}(y_i))$ ,  $i = 1, \dots, n$ . Then all the other steps of the original MCMC algorithm proceeds with the current state of the latent variables  $y_i^*$  ( $i = 1, \dots, n$ ) instead of the  $y_i$  ( $i = 1, \dots, n$ ).
- Spatio-temporal setting: In such a setting, we may specify the covariance function  $\sigma_C^2 \mathcal{C}_\phi(\mathbf{x}, \mathbf{x}') = \sigma_C^2 \mathcal{C}_{\phi_1}(\underline{\mathbf{x}}, \underline{\mathbf{x}}') \mathcal{C}_{\phi_2}(\mathbf{s}, t; \mathbf{s}', t')$ , given covariates  $\underline{\mathbf{x}}$ , spatial locations  $\mathbf{s} \in \mathbb{R}^K$ , and time  $t \in \mathbb{R}$ , where  $\mathcal{C}_{\phi_2}(\cdot, \cdot)$  denotes a correlation function for non-separable space and time effects (Gneiting & Guttorp, 2010). For example, the covariance function:

$$\begin{aligned} \sigma_C^2 \mathcal{C}_\phi(\mathbf{x}, \mathbf{x}') &= \sigma_C^2 \mathcal{C}_{\phi_1}(\underline{\mathbf{x}}, \underline{\mathbf{x}}') \mathcal{C}_{\phi_2}(\mathbf{s}, t; \mathbf{s}', t') \\ &= \sigma_C^2 \exp(-.5 \|\underline{\mathbf{x}} - \underline{\mathbf{x}}'\|^2) \exp(-.5 \{\|\mathbf{s} - \mathbf{s}'\|^2/2 + \|t - t'\|^2/2\}). \end{aligned}$$

- Estimating  $D_\tau(m)$ : For a given Bayesian model  $m$ , the estimate of the criterion  $D(m)$  is obtained by  $\widehat{D}(m) = \frac{1}{S} \sum_{i=1}^n \{y_i - y_i^{\text{pred}(s)}\}^2$ , given posterior predictive samples  $\{\{(y_i^{\text{pred}(s)}|\mathbf{x}_i, m)\}_{i=1}^n\}_{s=1}^S$ .

## Acknowledgements

This research is supported by National Science Foundation research grant SES-1156372, from the program in Methodology, Measurement, and Statistics. This paper was presented, in part, at the invited session on Bayesian nonparametrics, of the ERCIM Working Group conference on Computing and Statistics, Oviedo, Spain, December, 1-3, 2012

<b>Generating</b>	$D(m)$		
<b>Model:</b>	<b>GP</b>	<b>Indep</b>	<b>BART</b>
2-dimensional ( $n = 225$ )	5.0 ( $\pm 1$ )	75.2 ( $\pm 5$ )	48.7 ( $\pm 6$ )
10-dimensional ( $n = 100$ )	2.9 ( $\pm 3$ )	577.8 ( $\pm 7.4$ )	140.2 ( $\pm 3.1$ )
Multimodal ( $n = 100$ )	2.3 ( $\pm 1$ )	18.0 ( $\pm 4$ )	2691.0 ( $\pm 9.1$ )
Multimodal ( $n = 225$ )	27.6 ( $\pm 2.5$ )	316.9 ( $\pm 7.4$ )	6415.4 ( $\pm 11.4$ )

Table 1: Results of the Simulation Study. Predictive accuracy of the GP-based regression model, versus the independence model. (Each number in parentheses gives the corresponding 95 percent MCCI.)

## FIGURE CAPTIONS

Figure 1. The mixture weights  $\omega_j(\mathbf{x})$  and corresponding predictive density  $f(y|\mathbf{x})$  of the model. The figure assumes  $\eta(x_1) = -.30$ ,  $\eta(x_2) = .21$ ,  $\eta(x_3) = 4.8$ , and the covariance function  $\sigma(x, x') = \sigma_C^2 \exp(-.5||x - x'||^2)$ .

Figure 2. For the GP model, the posterior predictive mean of the change score (solid line) plus/minus 2 times the posterior predictive variance (dashed lines).

Figure 3. For the GP model, the posterior predictive density estimates, given a range of pretest scores, and conditional on  $uic=1$  and  $lmt140=1$ .





